opentext™

Maximizing document review efficiency with rapid analytic investigative review

opentext™

Contents

Executive summary

Legal teams review and analyze documents and data in a variety of contexts, from outgoing review for production in litigation to antitrust requests, third-party subpoenas and subject rights requests (SRRs).

Exigent deadlines often make it impossible to review every document, and risk tolerance may well obviate the need to review every document before disposition. Even when a comprehensive review is necessary prior to disposition, that same level of review may not be needed throughout the entire workflow. Each situation is unique. However, most teams apply a “one size fits all” methodology using either a linear review or a technology-assisted review, in which reviewers must review and code each document.

Rapid analytic investigative review (RAIR) is an alternative document review methodology that can be used in place of, or in conjunction with, traditional document review. RAIR can often be the most efficient and effective review technique in virtually any production scenario.

This paper explores RAIR methodology and examines its efficacy and efficiency by comparing it with more traditional managed document review in a variety of real-world use case scenarios.

opentext™

Introduction

Most modern document review scenarios—whether a litigation production review, a third-party subpoena review or even an internal or regulatory investigation—rely on traditional managed document review workflows. Regardless of whether the review is a linear review or a technology-assisted review, most documents are batched to teams of reviewers who apply independent judgments to each individual document before ultimate disposition.

While there are certainly situations that demand an eyes-on review before, for example, being produced to an opposing party, such is not always the case. Exigent deadlines often make it impossible to review every document. Risk tolerance may well obviate the need to review every document before disposition. Even when comprehensive review is necessary, the same level of scrutiny may not be needed throughout the entire workflow.

As a result, sophisticated review teams are increasingly incorporating rapid analytic review techniques, either as part of traditional review workflows or as a complete replacement. Relying on advanced analytics to aggregate groups of documents for collective assessment, these techniques are usually less expensive, less time-consuming and more accurate than even technology-assisted document reviews.

This position paper outlines the characteristics of a rapid analytic investigative review (RAIR) and details precisely how and why this technique can be more effective. The paper then uses several modern document review scenarios to demonstrate the relative strengths and weaknesses of various review techniques, both alone and in combination.

1. Understanding a rapid analytic investigative review

RAIR focuses on using advanced analytics to locate characteristically similar sets of documents that can confidently be managed as a group for purposes of ultimate disposition (e.g., production). Even though only a fraction of the documents in the review set are actually directly and independently reviewed, team structure and interdependent process workflows ensure maximum consistency, superior accuracy and defensibility.

Aggregating characteristically similar documents

The essence of RAIR is the aggregation of documents that are substantively similar (from the perspective of the ultimate decision), in such a way that the entire amalgamated set can be subject to a single decision.

The first step in the process is the same as it is in any managed document review—defining the decision-making criteria, whether responsiveness, privilege, issue classification or something else. This will define the ultimate objective and the touchstones by which review decisions will be assessed. Then documents are reviewed collectively to determine whether any analytics illuminate patterns that are more likely to fit into one or the other category. There are any number of ways to start this analysis.

Sender domains and subscription terms (e.g., “unsubscribe”) can often be used to quickly aggregate large sets of documents that are likely to be unnecessary without any further refinement. On the other end of the spectrum, search terms derived from review criteria will typically provide a substantial starting point for locating documents amenable to positive classification, with or without refinement. Those documents will, in turn, provide valuable insights into other search terms and analytic criteria, spring boarding into an even greater, ever-increasing set of analytic patterns that enable data aggregation and document classification.

For example, communication patterns evinced by emails identify those individuals whose documents should be aggregated for further analysis, refinement and assessment leading to classification. Similarly, the original file names and folder names associated with electronic documents (other than emails) establish organizational conventions that often lead to apparent aggregation and classification criteria.

Assessing the collected groups

As groups of documents are assembled, the next step in the process is to assess each group to confirm that the documents are amenable to bulk classification. This is done by taking a random sample having at least a 95% confidence level and ±5% confidence interval from each group (or, for smaller groups, using the entire set), and passing them on to a direct, managed review.

If review of the sample confirms that the entire group is consistent, all documents in the group are appropriately classified. If any of the documents in the sample are inconsistent with a unitary classification, the inconsistent documents are further evaluated to determine whether there are any inherent characteristics that suggest the need to further subdivide the group. If so, the group is subdivided and the sampling process is repeated for both document sets.

This process of aggregating, sampling and classifying continues until there are no analytic similarities between the documents remaining to be classified. At that point, the remaining documents are reviewed and classified individually.

Non-text documents (e.g., images, A/V files, etc.) often do not have sufficiently informative metadata to permit aggregation. To improve efficiency in the review of these non-text documents, they are typically de-duplicated based on their hash values with only a single instance of each document being reviewed and coded or classified. The results of this review and classification are then propagated to all hash duplicates.

Managing the process to optimize results

During a RAIR, the review team and workflows should be structured to take full advantage of the aggregation to deliver enhanced efficiency, accuracy and defensibility in comparison with other review techniques.

Responsibility for aggregation and initial classification is typically limited to one, or at most two, team lead individuals. This provides two qualitative benefits not available with any other review technique. First, the team lead(s) become intimately familiar with the overall characteristics and context, focusing their perspective on the review as a whole. Decisions reflect a comprehensive understanding of the implications on the ultimate objectives of the review and production. Second, a team lead also exercises well-informed quality control by virtue of making the initial aggregation decision and considering the decisions made during the sample review.

As groups of documents that may need privilege review are identified in classification, a dedicated privilege reviewer makes privilege decisions in coordination with the team lead. The ultimate result is a more consistent classification with respect to both individual coding decisions and the overall approach. Since the direct review of sample sets derives from documents aggregated based upon characteristic similarity, there is an inherent consistency at this second level, one that is not available with other review methods.

Every other managed document review technique, including technology-assisted review (TAR), relies upon batching documents to multiple reviewers. Whether through email threading, or even TAR prioritization, different reviewers will inevitably end up reviewing batches containing characteristically similar documents. The application of individual, independent perspectives will undoubtedly drive inconsistency.

The benefits of RAIR in depth

Speed and efficiency

Aggregation takes place much faster than an individual document review and far fewer documents are reviewed with RAIR than any other review technique, including continuous active learning TAR.

Additionally, accuracy of a RAIR will likely exceed that of any other review technique, considering both recall[1] and precision.[2] Most other review techniques simply cease once recall has exceeded 70% to 80% recall. There is no stopping point for a RAIR. Since the entire review set is typically assessed, it is not unusual to see recall levels exceeding 90%.

Given that every aggregated document set is sampled to 95% ±5%, together with the inherent consistency running all review through a single reviewer, the precision is typically similarly high, even relative to other review techniques.

Defensibility

RAIR's inherent defensibility derives from incorporating recording throughout the entire process and validation. As documents are aggregated, they are foldered and the basis for aggregation (e.g., search terms used, metadata analyzed, etc.) is recorded, providing a roadmap from the aggregation concept to the specific documents that were collected. Random samples used for assessment are also foldered to maintain the integrity of the analytic process and all associated reviews and classification decisions.

As with every other technique, the results of a RAIR can also be validated by appropriate sampling to confirm effectiveness. The proliferation of samples across small segments of the review collection ensures granularity and consistency with overall objectives. Combining this approach with dedicated, constant and consistent privilege review provides maximum protection from any unintentional disclosure or production. The incorporation of standard validation techniques ensures that any production objectives are met.

2. Incorporating RAIR techniques into traditional document review scenarios

RAIR can be used in virtually any document review scenario, either with or as a substitute for traditional review techniques. While every technique has strengths and weaknesses, RAIR will almost always provide benefits, even when it is necessary to review every document before production to another party.

Review for production in litigation

In considering the application of a RAIR in the context of a review for production for litigation, the primary question is whether it is necessary to independently review every produced document before it is produced.

If the answer is "no," then a RAIR will generally be more efficient and effective than almost every other traditional document review technique. Even if the answer is "yes," incorporating rapid analytic investigative techniques into a traditional review will save substantial time and costs.

1 Recall measures how many of the relevant documents in a collection have actually been found. For example, a 60% recall rate means that 60 percent of all relevant documents in a collection have been found and 40% have been overlooked.

2 Precision measures how many of the documents retrieved are relevant. For example, a 75% precision rate means that 75% of the documents retrieved are relevant, while 25% of those documents have been misidentified as relevant.

Consider a not-atypical document collection of 400,000 documents that is 15% rich.[3] At that level, 60,000 responsive documents are available for review and production. Although production would also encompass family members of the responsive documents, family production is inherent in the ultimate production of documents regardless of review technique. Accordingly, the review of family members does not play a significant role in evaluating the relative efficiency of alternative review techniques – the focus in on the effort required to identify responsive documents.

Statistically speaking, it is necessary to review at least 80% of the collection to achieve a reasonable, (typically) judicially acceptable recall level of 80%. Linear review of the entire collection is the norm – that means that at least 320,000 documents will be reviewed over the course of even a limited linear review for production. A RAIR will be more efficient.

Efficiency in depth

RAIR will typically be more efficient than modern technology-assisted reviews when it is not necessary to independently review every document before it is produced. The overall efficiency of technology-assisted review, based on the continuous active learning protocol,[4] can be measured by the average number of documents that must be reviewed to find one responsive document.

TAR efficiency generally increases as richness increases—for example, it may be necessary to review roughly six or seven documents to find a single responsive document when the collection is only 1% rich, while it may only be necessary to review two documents to find a responsive document when the collection is 15% rich (not accounting for non-text image documents, which go through a separate workflow). Thus a continuous active learning TAR would require the review of roughly 96,000 documents to attain an 80% recall level.

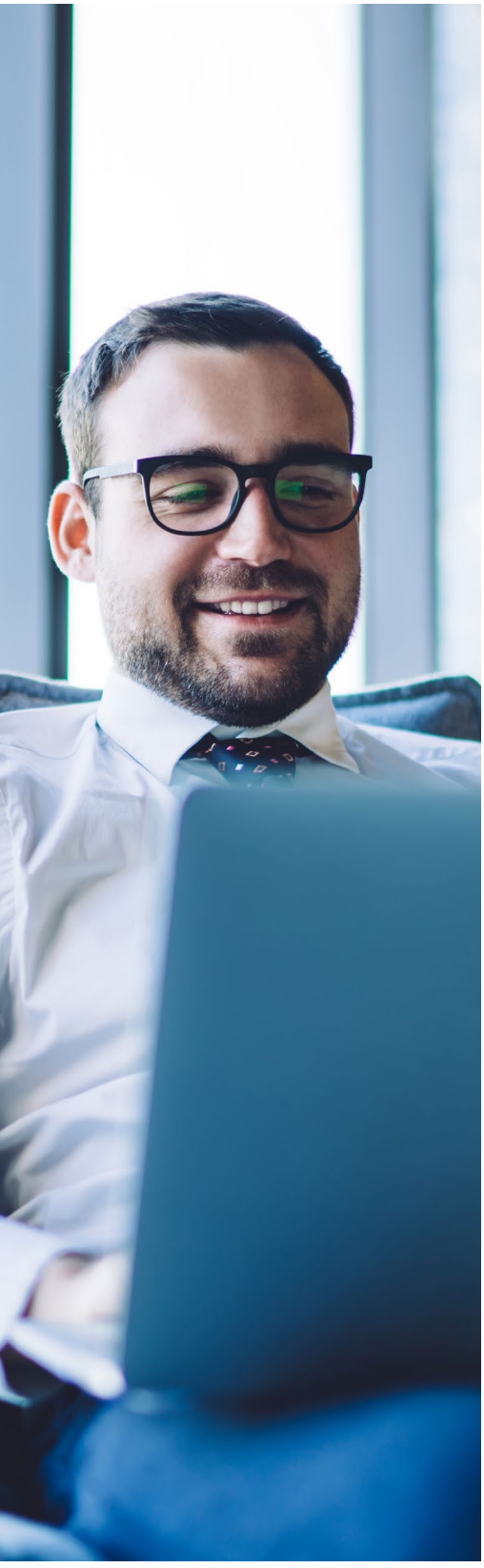
The ultimate efficiency of a RAIR depends to a large extent on the homogeneity of the collection—i.e., the extent to which documents can be aggregated into sets of characteristically similar documents. However, it is generally the case that documents in a given litigation share many of the same contextual characteristics, facilitating relatively effective aggregation. It is easy, then, to see how a RAIR can be more efficient than either a linear review or a TAR based on continuous active learning.

For example, even if 20% of the responsive documents were not amenable to aggregation, only 12,000 documents would require independent review. The remainder of the collection (48,000 responsive documents) would be assessed based on samples of aggregated sets of likely responsive documents. Even if a conservative average of no more than 1,000 documents could be aggregated into sets for sampling and review, that would generally lead to the review of 48 samples of roughly 385 documents (at 95% ±5%), or another 18,480 total documents reviewed—to assess the entire set of responsive documents.

That equates to an independent review of only 30,480 documents using RAIR, only 32% the size of a TAR based on continuous active learning (and obviously less than 10% of a linear review), while likely achieving an even greater recall level.

3 Richness is a metric of how many relevant documents exist in the dataset.

4 There are differences between earlier "TAR 1.0" systems, which are based on "simple learning": simple passive learning (SPL) and simple active learning (SAL) and "TAR 2.0" systems based on continuous active learning. While there are other differences between the approaches, SPL uses randomly-selected documents for training. The word passive refers to the fact that neither the algorithm nor the SME identifies documents for further training. SAL relies on the algorithm to select the documents used for training. Typically, the algorithm selects documents it is least sure about, in an attempt to identify the boundary between relevant and non-relevant documents. Continuous active learning continually learns as the review progresses, and regularly reranks the document population based on what it is has learned. As a result, the algorithm gets smarter and reaches its goal sooner and reviews fewer documents than would otherwise be the case with one-time training.



The insight factor

While RAIR efficiency may not match a simple learning TAR based purely on the number of documents subject to independent review, the qualitative benefits derived from RAIR may well exceed the cost of additional review.

A simple learning TAR review requires independent review of only somewhere between 5,000 and 10,000 documents, making an efficiency comparison to RAIR obvious. However, the level of knowledge derived from a simple learning TAR review pales in comparison to a RAIR, and the reality of over-production in a simple learning TAR presents a compelling reason for using RAIR instead.

Typically, a simple learning TAR relies on a control set representative of the collection, and training using uncertainty sampling—the review and assessment of documents that are neither clearly responsive nor clearly non-responsive. As a result, the subject matter expert (SME) conducting the review is only modestly exposed to the critical substance of the documents.

Although the control set spans the entire collection, the limited number of documents in the sample do not provide substantial insight into its vagaries. Most of the training review is devoted to only those documents that have tangential relevance. Consequently, review and production does not engender much knowledge about the substance, content or scope of the collection.

In contrast, a RAIR relies on generating significant insight into the document collection. The SME is tasked with locating and aggregating responsive documents, using every available analytic technique. As a result, the SME becomes exposed to, and even sensitive to, the substance of every set of documents, and the full scope of the collection. So the level of knowledge far exceeds what can be gleaned in a simple learning TAR.

Greater control

A RAIR also provides much more control over production parameters, avoiding the potential for the (sometimes damaging) over-production inherent in a TAR 1.0. Even the best simple learning TAR is typically no more than 60% precise—meaning that 40% of the documents produced using TAR based on simple learning are non-responsive.

Compounding that over-production is the fact that very little is known about the non-responsive documents being produced. They are certainly not reviewed, since production without an independent review is the essence of a simple learning TAR. Perhaps even more importantly, a simple learning TAR typically relies on an automated privilege/confidentiality screen that does not contemplate the independent assessment on many, if any, produced documents.

By comparison, the only documents produced in a RAIR review are documents that are very likely (based on SME aggregation and sample review and assessment) to be responsive. Accordingly, precision will be significantly greater than 60%—typically in excess of 90%. Knowledge of the documents being produced is a direct and inevitable result of the analytical aggregation of characteristically similar documents. Ultimately, while only a fraction of the collection is reviewed using RAIR, there is a reasonable understanding of the substance of every set of documents that is being produced.

Moreover, RAIR incorporates a dedicated privilege review that uses analytics to streamline decision-making. In combination with overall assessment by the SME, the RAIR approach provides a much more sophisticated and effective protection against disclosure of privileged or confidential documents.

Even when it is necessary to review every document before production, a RAIR can improve the efficiency of even the most efficient continuous active learning TAR. As noted above, a continuous active learning review of a collection that is 15% rich will require the average review of roughly two documents to find a single responsive document—a total of 96,000 documents in our example. Using a RAIR review to aggregate sets of responsive documents will, in turn, eliminate the review of some number of the associated non-responsive documents that would need to be reviewed using TAR based on continuous active learning.

So, for example, a RAIR that aggregates 20,000 responsive documents for further TAR review based on continuous active learning would, (based on a two-to-one efficiency), eliminate the need to review as many as 20,000 non-responsive documents. Thus, combining RAIR and TAR based on continuous active learning could reduce independent document review to 76,000—a 21% reduction—even when every responsive document is reviewed before production.

Review of opposing party productions in litigation

Generally, the review of an opposing party production should be devoted to finding only the specific documents that constitute, or lead to, information or evidence that will be useful. A true focused investigation approach will be most appropriate and effective. However, when overall classification or categorization of an opposing party production is the objective, RAIR will be much more effective than any other review technique.

While review efficiency may be important to an opposing party production review, it is likely more important to garner as much insight as possible into the substance of the production; and ensure consistency in the classification of documents. In that case, RAIR is far more beneficial than other review techniques.

Rather than having a diverse review team making classification decisions, decision-making is restricted to one or two team leads that aggregate and sample similar documents for review by a small, focused team of reviewers. In making aggregation decisions and recording the basis for those decisions for defensibility purposes, the team lead needs to fully understand and make critical distinctions between the substantive characteristics of the diverse documents in the collection.

As a result, the team lead gains valuable insights into the scope of, and the various substantive topics addressed in, the document collection. Consolidating decision-making in that way, together with confirmation by a small review team that focuses separately on each aggregated set of documents, further ensures consistency in classification.

That same level of knowledge and consistency simply cannot be achieved using any other review technique. Although the magnitude of review may differ, both a linear review and TAR based on continuous active learning rely on document batches being passed to large review teams. As a result, knowledge of the substantive contents of the collection (to the extent that review is not myopically focused only on the independent characteristics of each document) is spread amongst the entire team—there is no single individual to assimilate information the way a team lead does in RAIR.

Similarly, the distribution of document batches compels independent assessment across the entire review team, obviously impairing the potential level of consistency attainable with a focused RAIR.

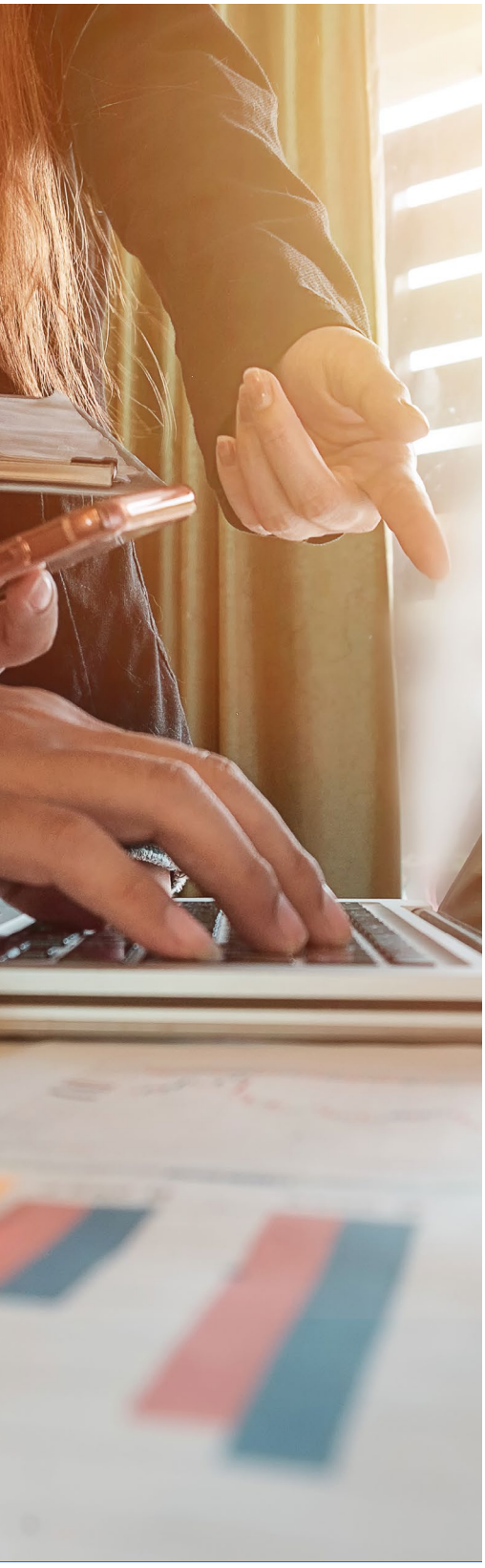
A simple learning TAR is not meant to develop any level of understanding about the substantive characteristics of a collection—training focuses only on those documents about which the TAR tool knows the least, leaving the vast majority of the collection completely unreviewed.

Third-party subpoenas

Third-party subpoenas present the quintessential opportunity for a RAIR. Since the producing party is not an active participant in the underlying litigation, there is no need to review every document before production. Minimizing cost is the driving factor, but limiting production of unnecessary documents is often a legitimate concern as well. Given those factors, there are only two appropriate choices:—TAR based on simple learning and RAIR.

Both of these techniques anticipate a limited review, without the independent inspection of every document being produced. A simple learning TAR will likely be slightly less costly and require minimal training (i.e., review), typically on the order of 5,000 to 10,000 documents.

A RAIR contemplates the review of one or two samples of every aggregated set of characteristically similar documents, typically at a 95%±5% level, or roughly 385 documents. By comparison then, a RAIR consisting of more than 13 to 26 aggregated sets of documents would likely be more expensive than an equivalent simple learning TAR. And, while generalizations are difficult, the increasing volume of review collections and the comparatively enhanced effectiveness a RAIR (exceeding the routine 80% recall target associated with TAR), mean that the RAIR sample review will likely exceed TAR simple learning training.



That said, a RAIR provides two qualitative benefits not available with a simple learning TAR. As discussed above, RAIR is typically much more precise. That means that far fewer non-responsive documents are unnecessarily produced, which is particularly critical for serial litigants or, frankly, in any context in which production needs to be circumspect. The critical assessment encompassed by the aggregation effort during a RAIR provides an insight into the contents of the production that simply cannot be attained through a TAR.

It is also possible to follow a simple learning TAR with a RAIR to minimize the number of non-responsive documents produced in response to a third-party subpoena. In that case, the RAIR would focus on swiftly eliminating non-responsive documents only—essentially the converse of a typical RAIR for responsiveness. Such an approach would improve overall precision but would not provide any additional insights into the substance of the production.

Second Requests (U.S. Hart Scott Rodino Antitrust Second Requests, EU Commission, UK CMA)

Second requests associated with mergers and acquisitions present perhaps even a more apt opportunity for a RAIR. Since it is usually not necessary to review every document before production to the agency, simple learning TAR and RAIR are viable review techniques—for the same reasons discussed as when responding to third-party subpoenas. Additionally, the tight deadlines often associated with these types of reviews virtually demand these methods' efficiencies.

Given the typical breadth of second requests and the sensitive commercial nature of the documents being requested and produced, it is even more critical to be careful and circumspect in production—and knowledgeable about the contents. For example, it would certainly be helpful (for planning purposes) to know well in advance whether the production suggests any inappropriate financial activities or recordkeeping, or some manner of improper anti-competitive conduct. These types of forewarning insights are only available with a RAIR.

RAIR also orchestrates the ultimate effective validation of a production to the agency in advance. While a simple learning TAR can only be validated once training and classification (as producible or not) have concluded, a RAIR is effectively validated throughout the course of the review.

The sample(s) of each of the aggregated document sets can roughly be considered stratified samples of a portion of the production. Since every sample relies on at least a 95%±5% statistical assessment (which is the same level of statistical accuracy required at validation), a RAIR virtually guarantees that the ultimate validation of the production will achieve similar levels.

SARs, SRRs, DSARs

Modern privacy regulations include the right of data subjects to demand certain information about the scope and content of personal data managed by corporate entities—variously referred to as Subject Rights Requests (SRRs), including Data Subject Access Requests (DSARs). These requests present yet another opportunity to leverage RAIR.

opentext™

Learn more

OpenText eDiscovery Services

The tight deadlines and minimal substantive review obligations associated with SRRs warrant efficient review techniques such as TAR based on simple learning[5] and RAIR. However, the nature of an SRR may not be wholly amenable to a TAR. SRRs depend primarily on the relationship between the document and the data subject, they are not particularly dependent on the content. Since a TAR is driven by assessment of substantive content, it may not be as effective.

RAIR, on the other hand, is an analytical assessment of documents, using any and all available analytics to aggregate characteristically similar documents. It does not suffer from the less-than-critical association between substantive content and producibility. RAIR can apply analytics to take advantage of any document feature tied to producibility to aggregate the appropriate documents. As a result, RAIR can be a very effective technique for locating documents that need to be produced in response to any subject access request.

RAIR also provides an effective screen against the production of non-responsive documents when, for example, documents contain personal information for other data subjects. Documents that are not producible can be identified and segregated using analytics in a RAIR to an extent that cannot be replicated using TAR (again, particularly because substantive content is not critical).

3. Conclusion

A rapid analytic investigative review can often be the most efficient and effective review technique in virtually any production scenario. RAIR relies on a small team to aggregate documents that are characteristically similar from the standpoint of the production obligation. Throughout the review, the RAIR team gains valuable insights into the substance and characteristics of the documents and continuously samples the aggregate sets to ensure validation and production at the highest levels of recall and precision.

Based on the analytics, aggregation and sampling, documents will be bulk coded for production purposes, and will be produced without necessarily looking directly at every document before production. As a result, RAIR can be extraordinarily efficient, while at the same time providing insight into the substance of the documents that cannot be gleaned using any similarly efficient technique (e.g., TAR).

About OpenText

OpenText is a trusted eDiscovery partner of law firms and organizations worldwide, including 15 of the world's 20 largest law firms, boutique firms and a majority of the world's largest tech companies. With a global footprint and deep expertise developing and optimizing proprietary eDiscovery technology with a suite of comprehensive services, OpenText helps customers navigate the challenges of modern litigation, investigations and regulatory response.

Connect with us:

- OpenText CEO Mark Barrenechea's blog
- Twitter | LinkedIn

5 A TAR review based on continuous active learning is unnecessary, since review of every document is not required before production to a data subject.

opentext.com/contact